A Sentence Similarity Measure Based on Conceptual Elements

Wendy Tan Wei Syn, Bong Chih How and Dayang Hanani Abang Ibrahim
Faculty Computer Science and Information Technology, Universiti Malaysia Sarawak, Sarawak, Malaysia.
wendytws@siswa.unimas.my

***Abstract*—There has always been a growing interest in sentence similarity measure for practical NLP tasks using various state-of-art NLP methods. Some of the widely used methods in measuring sentence similarity are lexical semantics, deep learning, neural networks, ontology, statistical models, graph based model and etc. Based on our findings, one of the main drawbacks in using these methods is not able to resolve word ambiguity where one word can have different interpretations in different sentences. In this paper, we present a sentence similarity measure by representing the sentences in conceptual elements to measure the semantic similarity between sentences. We used Microsoft Paraphrase Corpus (MSR) and Quora question pairs dataset to evaluate the performance. The study concludes that we were able to use conceptual elements to measure sentence similarity with the highest micro averaged precision of 0.71.**

***Index Terms*—Sentence Similarity Measure; Concept; FrameNet.**

I. INTRODUCTION

Most of the sentence similarity measures derived from word's similarity, co-occurrence, word order, N-gram, synonym, antonym, and etc. However, two sentences that have different structure or even overlapped words can be semantically similar per se. For example *"I feel sad."* and *"My mood is down."*. While two sentences that shared 80% of identical words can be dissimilar. For example, *"I am Andy"*. and *"I am sad."*. The sentences above obviously proved that bag of words (BOG) method has removed lots of detail yet less effective when the sentences contain ambiguous words conveying different meaning under different contexts. In order to find similar meaning sentences, hence semantically similar sentences, we need to go beyond word usage and sentence structure where a model can be trained to understand the concepts in the sentences.

A concept can be defined as "*a perceived regularity in events or objects, or records of events or objects, designated by a label"* [1]. Concepts is abstract. It is the mental representation of classes of things [2]. Concepts are represented with words or phrases. For example, the phrase of *"saves time"* represent the concept of efficiency.

When we want to understand a discussion, we are trying to grasp concepts using our background knowledge so that we can comprehend the statement. Here, concepts connect our past experience with the current interaction with the world [2]. Each concept is connected to one and another. We often discuss about common concept in our daily conversation. Throughout the conversation, we are using our own words when explaining something. For example, we might use different word such as *"wonderful!"*, *"fantastic!"*, but we are still conversing on the same concept which is expressing our feelings towards something.

In order to comprehence a sentence, besides trying to understand the each word's meaning, we have to capture the overall concept of the sentence as well, which are made up of words. As we know human have the ability to use memory as the inventory to structuring, classifying, and interpreting experiences [2], each particular word are associated in memory with particular frames (concept elements) [2]. For example, words such as *"buy"*, *"sell"*, *"pay"* able to activate the commercial event scenario in someone brain. Therefore it is crucial that in understanding a word's meaning requires knowing the whole scenario [2]. We might use the same word but referring to different other frames.

The same goes to when we want to identify if two sentences are similar, we should look at the similarity of concept besides overlapping words, syntactic similarity, or whether the sentence having the same subject-verb–object (SVO) or semantic role labelling (SRL). There are always possibilities that we might misunderstand the meaning if the above syntactic features are ambiguous. This illustrates the importance to focus on capturing the concept of a sentence in order to measure sentence similarity.

When we looking at the concepts, finding similar sentences mean finding sentences that are conceptually similar. When we represent the sentence as a concept, the words are categorized under a common concept which could help in sentence similarity measure. For example, *"This phone is easy to use"* and *"This phone is difficult to learn"*, the concept that we intend to capture for both of the sentences is the difficulty in using something. Throughout this paper, we will discuss on how to use conceptual elements in sentence to measure their similarity.

II. PROBLEMS

Recently, Google offshoot Jigsaw released a machine-learning-based service called Perspective which can be used to identify toxical comments to ensure the safety of Internet [3]. Perspective was trained from thousands of comments and was reported that the system tends to *"sensitized to particular words and phrases but not the meanings"* [3]. This clearly showed that the current AI approaches in understanding meaning in text remains a challenging issue especially when dealing with ambiguous scenarios.

Based on one of our experiment in using a computational model with the implementation of Latent Semantic Analysis (LSA) by Laudauer et. al. [4], we found out that in some cases the model failed to find related sentences which caused wrong classification. For example, one of the hiccup we run into was the following sentences *"Not to contradict myself, while Me functioned properly 80 percent of the time on my machine,*

there were times when it would continually crash upon loading or it would just not load." was reported to be similar with the following sentence, *"Shopping on the web saves time."*. This showed that LSA found a sentence that is similar to the concept of time, however, the actual focus of the first sentence was on "*crashing*" not "*saving time*". The model performed reasonably well in information retrieval but might not be able to differentiate words' meaning in certain context which caused it to retrieve irrelevant sentences. Somehow, by looking at the architecture of LSA, it used words co-occurrence to capture the meaning of sentences and similar meaning words tends to be positioned closer in the semantic vector space. This can be less flexible to fit in with wide domain words because the semantic model is not able to interpret different kind of sentences based on the context of training data. On the other hand, the semantic space was formed from words' patterns. For example, generally we know that the words *"happy"* and *"fun"* are similar. Let say we have a sentence, *S1: "I feel happy for this phone."*. When we used sentence similarity measure, our goal is to find any similar sentence to S1 that contain related feelings represented by significant word such as *"sad"*, *"fun"* instead of sentences that contain overlapping or unimportant related words such as *"system"*, *"software"* and etc. Which mean the meaning derification part can be ambiguous when there are several ambiguous words exist in the sentences.

For human, we know that the two sentences: *"This phone makes me happy."* and *"The system is fun."* are similar because we were focused on the main keywords *"happy"* and *"fun"* that represent pleasant feelings. In most of the cases, we are not really emphasizing on what is the object involved, what we are more interested to know is the feelings towards something. This also means that we need a model to automatically identify the main concept in a sentence in order to infer if the two sentences are similar.

As a first initiative to solve this problem, we proposed a sentence similarity measure based on conceptual elements. We derived the sentences' concepts through FrameNet's conceptual elements, a theory that based on frame semantics. We believe that every sentence can be represented with conceptual elements which together infer to a main concept. In order to know if two sentence are similar, we intend to investigate from the perspective of similarity in concepts so that we can resolve words' ambiguity if they are referred to different contexts. From there, we can start to use the conceptual elements to compare sentences similarity which we will discuss in details in the followings.

III. METHOD

In this study, we represent sentences in concept by using FrameNet. We will describe our proposed method.

A. FrameNet

Framenet is based on a theory of meaning called frame semantics [5]. The central idea of Frame Semantics is word meanings must be described in relation to semantic frames – schematic representations of the conceptual structures and patterns of beliefs, practices, institutions, images, etc. that provide a foundation for meaningful interaction in a given speech community [6]. According to Baker [5], the meaning of most words can best be understood on the basis of semantic frame: a description of a type of event, relation, or entity and the participants in it. For example, the concept of cooking typically involves a person doing the cooking (Cook), the food that is to be cooked (Food), something to hold the food while cooking (Container) and a source of heat (Heating_instrument). In the FrameNet project, this is represented as a frame called Apply_heat, and the Cook, Food, Heating_instrument and Container are called frame elements (FEs). Words that evoke this frame, such as fry, bake, boil, and broil, are called lexical units (LUs) of the Apply_heat frame [5].

FrameNet is different with WordNet where FrameNet contain semantic roles and evoke frames (type of events, relation, or entity). WordNet on the other hand clusters partially synonymous words in "synset" form. The table below showed how FrameNet and WordNet distinguishes the word "curiousity" in different senses [7].

FrameNet	WordNet
Frame: Typicality **Definition:** Unorthodox or unexpected	**Syn. Set:** Curious, funny, odd, peculiar, etc. **Definition:** Beyond or deviating from the usual or expected
Frame: Mental_stimulus_exp_focus **Definition:** Interested or inquisitive (about something)	**Syn. Set:** Curious **Definition:** Eager to investigate and learn or learn more; having curiosity aroused; eagerly interested in learning more
Frame: Mental_property **Definition:** Driven to investigate and learn	

Figure 1: FrameNet vs WordNet in distinguishes the word "curiosity" in different senses

Based on Figure 1, FrameNet ragards *curiosity* as a character trait and also a mental state while WordNet is storing the word's synonym. Therefore, FrameNet can be suitable to be used in representing concept for a sentence. If we able to identify frames in different sentences, then we could factor in the concept to derive semantic similarity between sentences. In order to identify whether two sentences are similar, we are not only need to understand the meaning of each sentence, but also need to differentiate their meaning.

B. Using Conceptual Elements in Sentence Similarity Measure

By adopting FrameNet in our study, we can represent sentences with conceptual elements. In the following, we converted few sentences into conceptual elements for illustration.

Table 1
Sentences and Conceptual Elements

Sentences	Conceptual elements (Frames from FrameNet)
1. I like using this website.	Experiencer_focus- Using
2. Using the Internet TV is enjoyable.	Using- Stimulus_focus
3. I enjoy using the Web.	Experiencer_focus Using- Network
4. Using the Web enhances my productivity.	Using-Network Cause_to_make_progress

Table 1 and shows the examples of concept elements derived from sentences. Table 2 shows the definition of the conceptual elements from Table 1. We can clearly see that sentence contain few conceptual elements. While for the semantic models such as LSA, the representation is derived from the corpus might not represent words' actual meanings under different contexts, if they are not covered the contexts. When we derived conceptual elements from a sentence, indirectly, we are categorize the same concepts to increase recall. By forming conceptual elements, we are not

eliminating words but fitting the words as informative frames to the sentences. In our works, we are comparing the conceptual elements instead of focusing on co-occurrence of the words. Our proposed model will be dealing with concepts instead of individual words in the sentence. The concepts are considered as the generalization of the original sentence.

Table 2
The Definition for Each Frame Stated in Table 1

Frames	Definitions
Experiencer_focus	The words in this frame describe an Experiencer's emotions with respect to some Content. A Reason for the emotion may also be expressed. Although the Content may refer to an actual, current state of affairs, quite often it refers to a general situation which causes the emotion.
Stimulus_focus	In this frame either a Stimulus brings about a particular emotion or experience in the Experiencer or saliently fails to bring about a particular experience. Some words indicate that the Stimulus is characterized by the experience it is likely to evoke in an Experiencer and for these, the Experiencer may rarely be present. There may also be a Degree to which the Stimulus affects the Experiencer and Circumstances under which the experience occurs. There may also be a Comparison_set to which the Stimulus is compared and a Parameter that indicates the area in which the Stimulus has its effect.
Cause_to_make_ progress	An Agent works on a Project so that it reaches a more advanced and desirable state.

In Table 3, the two semantically dissimilar sentences *"It is important to make shopping easy"* and *"It is important to minimize payment time."* are contextually different even though they shared a few overlapping words. The conceptual elements on the other hand can differentiate them. Besides that, we also believe that by forming conceptual elements, we are reducing the number of tokens in the sentences yet preserving the meaning. Let say we have another similar sentence S3 which is similar to S1. The semantic model does not need to infer *"easy"* is synonymous to *"less tough"* because they fit into the same concept elements: *"difficulty"*.

Table 3
Sentences and example of concept elements adapted from SEMAFOR tool

Sentences	Conceptual elements (Frames from FrameNet)
S1: It is important to make shopping easy.	Importance- Causation- Shopping- Difficulty
S2: It is important to minimize payment time.	Importance -Commerce_pay- Measure_duration
S3: It is important to make shopping less tough.	Importance –Causation- Shopping - Difficulty

IV. EVALUATION DATA

In order to evaluate our proposed sentence similarity measure, we adapted Microsoft Paraphrase Corpus (MSR) dataset [8]. We assume that two paraphrased sentences share similar concept. Thus, our goal here is to evaluate if our proposed measure is able to measure two paraphrased sentences as similar. MSR comprises 5801 candidate paraphrase sentences pairs which adapted from Web news sources. The sentence pairs are annotated by human judges. The dataset have been split into a training set with 4076 examples and a test set with 1725 examples. We also facilitated Quora duplicate question pairs (12339 question pairs with train set of 9871 sentences and test set of 6468 sentences) from Kaggle competition [9] for evaluation. We assume that duplicate questions pairs share similar meaning as well. Hence, MSR and Quora dataset will act as the gold standard to evaluate our proposed similarity measure.

V. EXPERIMENTS

Below are the steps involved in the experiments:

1. We converted all paraphrase and sentence pairs from our gold standard (both training and testing set) into frames using SEMAFOR parser [10] which based on FrameNet.
2. We then use Support Vector Classifier (SVM) (Parameters: C=1000000, gamma=10.0, kernel: RBF) to train a model based on the training set and use it to predict the category (1 as paraphrase, 0 as not paraphrase) for the testing set. We experimented with different features which derived from the generated frames to train and test the SVM. Frames refer to the conceptual frames obtained from Semafor parser. Take the above sentences, *S1 and S2* where *S1 and S2* are semantically identical paraphrase sentence pairs, we get the features as shown in Table 4.
3. Report the results in macro and micro averaged f-measure by comparing the category annotated in gold standard with SVM predicted category (1- paraphrase, 0- not paraphrase).
4. The same steps are repeated for Quora dataset with annotated category (1 as duplicated, 0 as not duplicated).

Table 4
Original sentence and represented frames

Original sentence	Represented frames
S1: "She was surrounded by about 50 women who regret having abortions.",	*S1: Locative_relation Relational_quantity Cardinal_numbers People Experiencer_focus Possession*
S2: "she was surrounded by about 50 women who have had abortions but now regret doing so"	*S2: Locative_relation Relational_quantity Cardinal_numbers People Possession Temporal_collocation Experiencer_focus Intentionally_act*

Table 5
Features for SVM training

Type of features	Features for SVM
Number of overlapping frames	5
Number of frames of s2 not in s1	3
Number of frames of S1 not in S2	0
Number of frames of s2 not in s1 & number of frames of S1 not in S2	3,0
Number of frames of s2 not in s1 & number of overlapping frames	3,5
Number of frames of S1 not in S2 &number of overlapping frames	0,5
Number of frames of s2 not in s1 & number of frames of S1 not in S2 & number of overlapping frames	3,0,5

Type of features	Features for SVM
Overlapping frames	Locative_relation Relational_quantity Cardinal_numbers People Experiencer_focus Possession

VI. RESULTS

Table 6 shows the result obtained. Based on the result, by considering the number of overlapping frames yielded the best result. The reason is that the paraphrase sentence pairs and Quora question pairs shared the meaning which can be captured by the conceptual frames. The effectiveness of applying conceptual frames can be observed after we converted the original raw sentences into conceptual frames. Besides that, by representing sentences in conceptual frames, it able to reduce noise which helps in inferring semantic meaning.

However for some sentences, the concept cannot be fully represented using the existing frames in FrameNet and caused loss of information. For example, the following sentence: *"The Calgary woman who is in her twenties donated blood on Aug 7"*. The conceptual elements obtained was only *"people giving"* which is insufficient to represent the meaning of this sentence. This insufficient information will produce wrong prediction. Besides that, we also found out we still face the problem of ambiguity and this problem cannot be fully resolved especially for sentences that have few common concept elements. For example, the following two sentences:

S1:" The company didn't detail the costs of the replacement and repairs"

S2: "But company officials expect the costs of the replacement work to run into the millions of dollars"

After converting them into conceptual elements:

*S1: "*Businesses Expensiveness *Take_place_of Self_motion"*

S2: "Businesses Leadership Expectation Expensiveness Take_place_of Working_on Leadership Quantity"

The two sentences above contain different meaning and concept. However, our model predicted them as similar sentence which we found out most probably was caused by the common conceptual elements *"Businesses Expensiveness"*. These common conceptual frames represented *"cost"* and *"business"* but by looking at the overall context, S1 and S2 mentioning different things which is not paraphrase. This showed that it is challenging to build the model will focus on the main concept instead of affected by high frequency of certain words. This highlighted the needs to have a model which can identify the main concept represented by keywords in a sentence. Nonetheless, our experiments proved that by adding conceptual elements into sentences can help in sentence similarity measure. This is similar to when someone does not understand your lengthy explanation on how difficult it was when using a product, you may explain it again by telling the underlying concept such as you are actually complaining about the difficulty in using that product so that one can roughly understand what you are trying to say.

Table 6
Macro and micro averaged precision for each experiment

Type of features	Macro averaged precision MSR dataset	Micro averaged precision MSR dataset	Macro averaged precision Quora dataset	Micro averaged precision Quora dataset
Original raw sentence (SVM classification)	0.48	0.68	0.44	0.64
Number of overlapping frames	0.61	0.69	0.39	0.62
Number of frames of s2 not in s1	0.42	0.67	0.38	0.62
Number of frames of S1 not in S2	0.44	0.66	0.38	0.62
Number of frames of s2 not in s1 & number of frames of S1 not in S2	0.51	0.67	0.48	0.61
Number of frames of s2 not in s1 & number of overlapping frames	0.59	0.68	0.56	0.62
Number of frames of S1 not in S2 & number of overlapping frames	0.59	0.68	0.58	0.63
Number of frames of s2 not in s1 & number of frames of S1 not in S2 & number of overlapping frames	0.58	0.67	0.58	0.63
Overlapping frames	0.57	0.71	0.51	0.63

VII. RELATED WORKS

Most of the current sentence similarity measure are based on lexical resources such as WordNet. Term-matching method usually failed to capture meanings. Corpus based method such as LSA uses statistical information from huge corpus to calculate sentence similarity. Other methods included graph based approaches [11], ontology [12] and deep learning [13]. Stayya et al. [14] introduced the concept-based similarity measure that used the vector of weighted terms to determine the similarity between the documents. They also implemented temporal—semantic similarity measure that included time entities to detect temporal sentences. Recski et. al. [11] presented the method of measuring semantic similarity of words using concept networks that using WordNet database and features extracted from concept dictionary to build a set of conceptual graphs. Another graph approaches by Zhu and Iglesias [15] was measuring the semantic similarity between concepts in Knowledge Graphs which used Information Content (IC) of concept to weight the shortest path length between concepts. Elavarasi et. al. [12] constructed an ontology to represent the knowledge as the set of concept to measure semantic distance. The shortest distance for each concept extracted from the ontograph is used to measure semantic weight. The work by Liebeck et. al. [13] had implemented three approaches to measure semantic textual similarity: 1) Use WordNet and word2vec to measure the overlapping between tokens in

sentences. 2) Train neural network model using the two features. 3) Implement surface–level similarity, context similarity and topical similarity.

VIII. Conclusion and Future Works

To conclude, we have presented a sentence similarity measure that uses conceptual elements. We focus on conceptual level semantic similarity instead of word patterns that can be ambiguous. Based on our result, we showed that by including conceptual elements into the sentences is able to improve the semantic similarity measurement performance. However, there are still a few issues needed to be solved such as ambiguous common concept and insufficient concept elements to represent the overall concept of the sentence.

As for future works, we can enhance the concept of a sentence by generating more topics using topic models such as Latent Dirichlet Allocation (LDA) [16]. We deduce that concept can be represented by topics, however we shall solve the main problem where the model should be able to capture the main concept in a sentence just like how human identify them. Besides that, we realized that not all the frames from FrameNet are applicable to sentences of different domain.

Acknowledgement

We would like to thank Universiti Malaysia Sarawak, MyPhD by KPT and MOHE who funded this project through grant ERGS/ICT07(01)/1018/2013(15). We also like to express our sincere gratitude to Kai Larsen, the Director of Human Behavior Project at University of Colorado Boulder for allowing us to use the behavior variables in the study.

References

[1] J. D. Novak and A. J. Cañas, "The theory underlying concept maps and how to construct and use them," *Florida Institute for Human and Machine Cognition*, 2008.

[2] C. J. Fillmore, "Frame semantics and the nature of language," *Annals of the New York Academy of Sciences*, vol. 280, no. 1, pp. 20-32, 1976.

[3] D. Auerbach, It's Easy to Slip Toxic Language Past Alphabet's Toxic-Comment Detector, 2017. Retrieved February 25, 2017, from https://www.technologyreview.com/s/603735/its-easy-to-slip-toxic-language-past-alphabets-toxic-comment-detector/#comments

[4] T. Laudauer, P. Foltz, and D. Laham, "Introduction to Latent Semantic Analysis," *Discourse Processes*, vol. 25, pp. 259-284, 1998.

[5] C. F. Baker, "FrameNet: A Knowledge Base for Natural Language Processing", *Proceedings of Frame Semantics in NLP: A Workshop in Honor of Chuck Fillmore (1929–2014)*, pp. 1-5, 2014.

[6] C. J. Fillmore, C. R. Johnson, and M. R. Petruck, "Background to framenet," *International journal of lexicography*, vol. 16, no. 3, pp. 235-250, 2003.

[7] The Blog of the International Computer Science Institute New Research: Aligning Lexical Resources, 2 May 2012. Retrieved February 23, 2017, from https://www.icsi.berkeley.edu/icsi/blog/aligning-lexical-resources

[8] Microsoft Research Paraphrase Corpus. http://research.microsoft.com/research/downloads/default.aspx

[9] Quora Question Pairs, https://www.kaggle.com/c/quora-question-pairs/data

[10] D. Das, D. Chen, A. F. Martins, N. Schneider, and N. A. Smith, "Frame-semantic parsing," *Computational linguistics*, vol. 40, no. 1, pp. 9-56, 2014.

[11] G. Recski, E. Iklódi, K. Pajkossy, and A. Kornai, "Measuring semantic similarity of words using concept networks," *ACL 2016*, p. 193, 2016.

[12] S. A. Elavarasi, J. Akilandeswari, and K. Menaga, "Ontology based Semantic Similarity Measure using Concept Weighting".

[13] M. Liebeck, P. Pollack, P. Modaresi, and S. Conrad, "HHU at SemEval-2016 Task 1: Multiple Approaches to Measuring Semantic Textual Similarity," *Proceedings of SemEval*, pp. 595-601, 2016.

[14] Satya P Kumar Somayajula et al, *(IJCSIT) International Journal of Computer Science and Information Technologies*, vol. 2, no. 4, pp. 1743-1746, 2011.

[15] G. Zhu and C. A. Iglesias, "Computing Semantic Similarity of Concepts in Knowledge Graphs," *IEEE Transactions on Knowledge and Data Engineering*, vol. 29, no. 1, pp. 72-85, 2017.

[16] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent dirichlet allocation," *Journal of machine Learning research*, vol. 3, pp. 993-1022, Jan. 2003.